

An Overview of the Palomar Transient Factory Pipeline and Archive at the Infrared Processing and Analysis Center

C. J. Grillmair, R. Laher, J. Surace, S. Mattingly, E. Hachopian,
E. Jackson, J. van Eyken, B. McCollum, S. Groom, W. Mi, and
H. Teplitz

*Infrared Processing and Analysis Center, M/S 100-22, California
Institute of Technology, CA, USA*

Abstract. The Palomar Transient Factory is conducting a wide-field, variable-cadence optical survey of the northern sky to detect transient, variable, and moving objects. As a member of the PTF collaboration, the Infrared Processing and Analysis Center has developed an image archive, a high-quality photometry pipeline, and a searchable database of detected astronomical sources. The system is capable of processing and storing 300 Gbytes of data per night over the course of the 5-year survey. With an expected total of ~ 20 billion rows, the table containing sources extracted from PTF images will be among the largest astronomical databases ever created. The survey is efficiently discovering transient sources from asteroids to supernovae, and will inform the development of future sky surveys like that of the Large Synoptic Survey Telescope.

1 Introduction

The Palomar Transient Factory (PTF) is a wide-angle, variable cadence, optical sky survey designed to detect all manner of transient objects, from asteroids to variable stars to supernovae. The cadences employed and the PTF key projects currently being pursued are summarized in Figure 1. The 5-day cadence survey is the largest in terms of sky area ($\approx 3\pi$ steradians) and is primarily aimed at finding Type II Supernovae. A more detailed description of the science cases for PTF is given by Rau et al. 2009.

The survey makes use of the Oschin 48-inch Schmidt telescope (P48) atop Palomar Mountain in California. The PTF camera is a $12K \times 8K$ CCD mosaic, originally built for the Canada-France-Hawaii telescope, now modified to provide faster detector readout and to operate in the more confined focal plane of the P48. The combination of telescope beam and camera provide 1.01 arcsecond per pixel sampling over a field area of 7.9 deg^2 . Typical survey exposure times are 60 seconds, yielding limiting magnitudes of $m_g \approx 21.4$ and $m_R \approx 20.6$. The large-area, 5-day cadence survey will be primarily conducted in the R-band. The instrument and data taking system are described more fully by Law et al. 2009.

PTF achieved first light in December of 2008 and began survey operations in May of 2009. Despite high atmospheric ash content (a result of large wildfires in the area) that precluded observing for extended periods during the last half of 2009, some 56,000 images ($\approx 10 \text{ TB}$) have been collected to date.

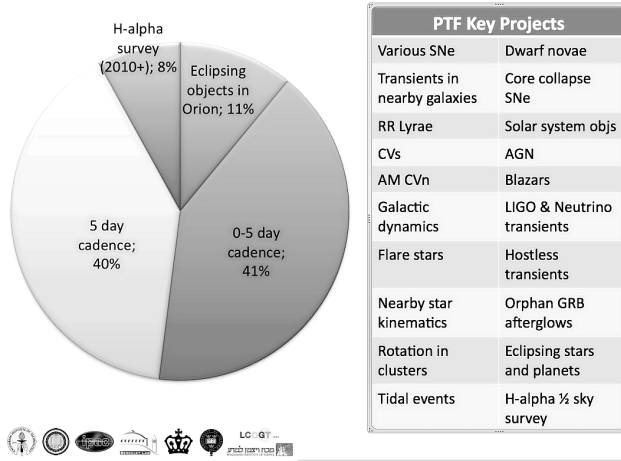


Figure 1. PTF cadences and the key projects they serve.

A data flow diagram for the PTF system as a whole is shown in Figure 2. By virtue of two distinct pipelines and a network of follow-up telescopes, PTF provides automatic, real-time transient classification and follow-up, as well as a database of every source detected in each frame. All images, both raw and processed, are stored on spinning disk and will be available for download.

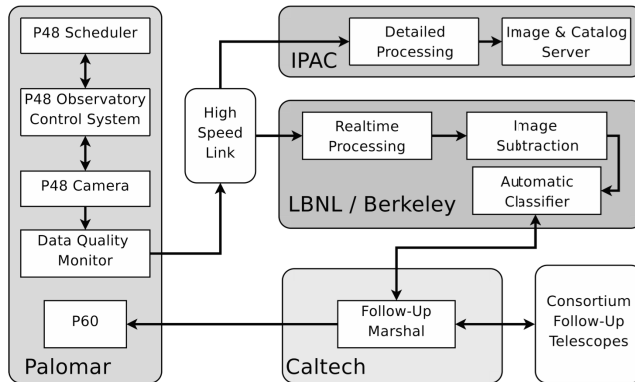


Figure 2. Data flow diagram for the PTF system as a whole.

In this contribution we describe the Infrared Processing and Analysis Center (IPAC) PTF pipeline and archive. A separate PTF pipeline at the Lawrence Berkeley National Laboratory uses image subtraction to detect transient sources in near real time and permit rapid classification and same-night follow-up. Conversely, the IPAC PTF pipeline operates on a less rapid time scale and is designed to achieve the best possible photometric accuracy consistent with hardware capabilities and data rate. Output from the IPAC pipeline is used to feed a large

database that can be queried through the NASA/IPAC Infrared Science Archive (IRSA).

2 IPAC-PTF Pipeline

In production mode, PTF generates ≈ 60 GB of raw data per night on average, with peak volumes approaching ≈ 100 GB on clear winter nights. The raw image data are sent from Mount Palomar to Caltech via fast microwave link and landline. The raw data are ingested daily at IPAC, which includes storage on spinning disk, registration of file location and MD5 checksum in the database, and verification against the nightly file manifest that is delivered with the data. Once all science and calibration files have been received, the photometry pipeline is initiated. All raw and processed images are stored on spinning disk and will be made available for public download after an 18 month proprietary period. A total of ≈ 300 TB of raw and processed image data will be stored over the course of the 5 year survey. This is approximately forty times the volume of image data stored by the 2MASS (Skrutskie et al. 2006) and SDSS (York et al. 2000) surveys. Pan-STARRS will produce data at ten times the rate of PTF, but nearly all of the image data will be discarded.

3 IPAC-PTF Computing Resources

The IPAC PTF pipeline hardware currently includes:

- 12 Sunfire x4150 8-core pipeline drones
- 2 Sunfire x4150 DBMS servers
- 1 Sunfire x4150 operations file server (software, sandboxes)
- 2 Nexsan SATAbeast 128 TB RAID-5 connected to IRSA file server for raw and processed data
- 1 Nexsan SATABeast 36 TB RAID-10 for operations file server and sandbox
- 1 Nexsan SATABeast 32 TB RAID-5 for database storage
- 4 Nexsan SATAbleds for primary file server (software, sandboxes)
- 5 Nexsan SATAbleds and 2 Sun 2540s for secondary database

An additional 60-100 TB of storage will be installed to complete the survey. Parallel processing on 11 multiprocessor Linux machines is employed to give processing throughput that meets the data-rate requirement, which is $5\times$ the real-time data acquisition rate. This requirement allows for processing nightly data as well as reprocessing older data, which will be necessary as the pipeline is refined and improved over time. Disk storage is configured as RAID-10 for data integrity. Both database and raw data are backed up on a weekly basis. In case of data corruption or loss, up to a week's worth of processing would have to be repeated in the worst case. A data flow diagram for the IPAC PTF pipeline is shown in Figure 3.

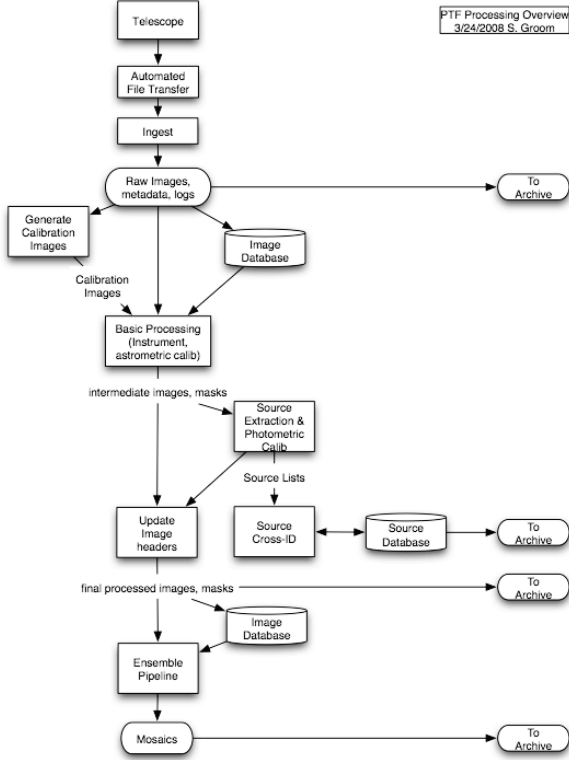


Figure 3. Data flow in the IPAC PTF pipeline.

4 Pipeline Software

Standard image processing modules (bias removal, flat fielding) are written in C. Where possible (e.g. when > 100 sky images have been taken), flat fields are generated each night by source-masking and median-combining all available science images. Though the PTF detectors have been in use for some time and suffer a considerable range in pixel sensitivities and cosmetic defects (with one detector currently being nonfunctional), these nightly flat fields are very well measured and reduce such variations to very low levels. Figure 4 shows the effect of this frame processing on a typical sky image.

Astrometry and photometry are carried out using community software, notably Astrometry.net (Lang et al. 2009), SExtractor (Bertin & Arnouts 1996), and Scamp (Bertin 2006). A high level flow diagram of the pipeline is shown in Figure 4. SExtractor is actually applied to each image a number of times in the pipeline to mask sources prior to flat field generation, measure the effective seeing, generate pixel weights and masks, and finally to generate photometry for all detected sources.

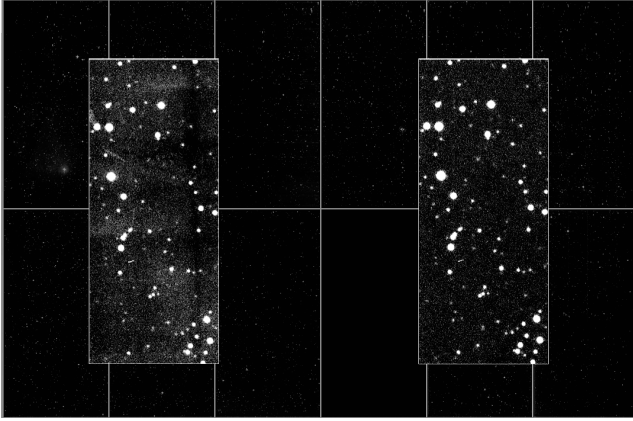


Figure 4. A comparison of typical sky image before and after frame processing.

The astrometric solution for each images is independently checked. The method employed consists of querying the database for appropriately bright and isolated 2MASS sources, matching them to sources extracted from the image, computing the RMS separation of the matches along the image axes, and requiring a minimum number of matches and both RMS values to be less than some threshold value. Currently a minimum of 20 matches and RMS values less than 1.5 arcsecond are required. Images not meeting these minimum requirements are thrown out.

The pipeline currently processes a night’s worth of data in 4 to 5 hours, ensuring that any necessary reprocessing of older data will be possible during normal operations without creating a backlog. Multi-threading is used by the pipeline software to achieve this processing rate.

5 Photometric Calibration

The goal of the IPAC PTF pipeline is to provide photometry accurate to 2% or better. An analysis of pipeline-generated flat fields indicates detector stability of $\approx 1\%$. Photometric calibration is carried out using stars observed in Sloan Digital Sky Survey (SDSS) fields. For all fields that fall within the SDSS footprint, a photometric solution for zeropoint and color term is computed by direct comparison with SDSS photometry. This is used to compute a nightly extinction coefficient, which is subsequently used to calibrate PTF photometry from images that fall outside the SDSS footprint.

6 Database

The PTF operations database is being run under open source PostgreSQL. The primary Sources table will contain astrometry and photometry of every source detection over the course of the survey. With between 20 and 40 billion rows,

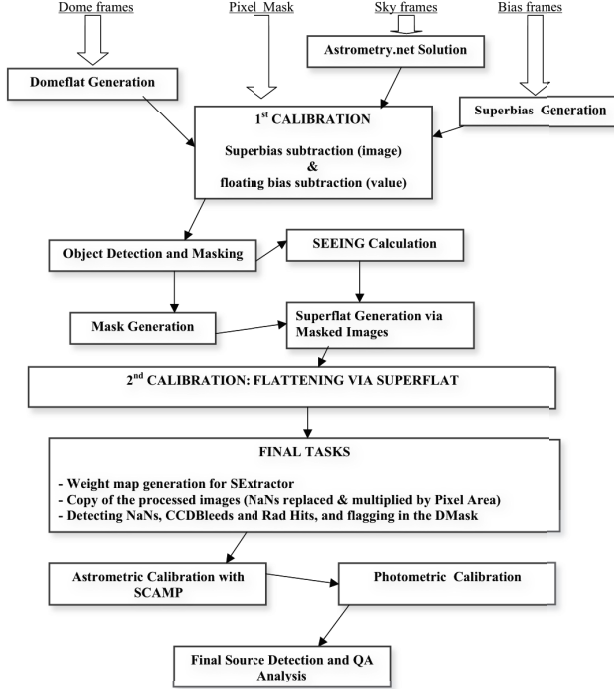


Figure 5. Schematic diagram of the IPAC PTF pipeline.

each with 69 photometric and astrometric parameters, this table will be among the largest ever created. Load testing under PostgreSQL has thus far shown no unexpected issues with regard to speed or scalability. Care has been taken in the database schema to ensure that database entries and indices are written to spinning disk via separate I/O channels.

A source association pipeline, run daily, will use each night’s detections to update a Merged Sources table. This Merged Sources table will provide time-series photometry for every detected source and is expected to be the primary science output of the IPAC PTF pipeline. We have developed an efficient method of source association that is done in parallel for independent declination zones and thus requires no inter-process communication.

7 IRSA Interface

The NASA/IPAC Infrared Science Archive (IRSA) provides easy access to PTF data, including both images and catalogs. Archive services for PTF include functionality for both the IPAC pipeline team, the PTF science team, and (eventually) the astronomical community. IRSA provides both raw and processed data, including full FITS frames, image cut-outs, and supermosaics. Image previews are augmented by 2MASS catalog overlays. The imaging data can be accessed

through temporal or spatial searches. Figures 6 and 7 show examples of the image preview feature.

IRSA provides access to PTF source tables through it’s search tool (“Gator”). The tables include both Sources (extracted from individual frames) and Merged Sources. A lookup functionality between the two tables will be added in the near future. Records can be accessed with a variety of search constraints for values in the table, including but not limited to position and time of observation. A current coverage map is shown in Figure 8. After several years of observation, the PTF Sources table will be among the largest tables served by IRSA (or any astronomical archive for that matter), rivaling even the WISE individual sources catalog.

PTF Nightly Summary for 2010-01-01	
Observation Info	Click Image for Preview
Date: 2010-01-01 04:10:45.509 Download: PTF201001011741_2_o_9293.fits Nid: 305 Expid: 51694 RA: 81.8241000 Airmass: 1.4263800 Dec: 1.8758000 Seeing: 4.3700000 Obsvr: Kulkarni Ptfpid: 40000 Moon RA: 108.0612170 Filter: R Moon Dec: 22.6301970 Exptime: 30.0000000 Phase: 354.9330000 Imgtype: object	
Date: 2010-01-01 04:43:13.809 Download: PTF201001011967_2_o_9318.fits Nid: 305 Expid: 51695 RA: 81.8247340 Airmass: 1.3122200 Dec: 1.8752000 Seeing: NaN Obsvr: Kulkarni Ptfpid: 40000 Moon RA: 108.3549970 Filter: R Moon Dec: 22.5958560 Exptime: 30.0000000 Phase: 354.6180000 Imgtype: object	
Date: 2010-01-01 07:01:10.659 Download: PTF201001012925_2_o_9424.fits Nid: 305 Expid: 51696 RA: 81.8215140 Airmass: 1.1826600 Dec: 1.8747000 Seeing: 2.4700000 Obsvr: Kulkarni Ptfpid: 40000 Moon RA: 109.4628830 Filter: R Moon Dec: 22.3916680 Exptime: 30.0000000 Phase: 353.2760000 Imgtype: object	
Date: 2010-01-01 09:59:45.508 Download: PTF201001014165_2_o_9561.fits Nid: 305 Expid: 51697 RA: 81.8243390 Airmass: 1.8906500 Dec: 1.8759000 Seeing: NaN Obsvr: Kulkarni Ptfpid: 40000 Moon RA: 110.7574830 Filter: R Moon Dec: 21.9519490 Exptime: 30.0000000 Phase: 351.5390000 Imgtype: object	

Figure 6. Screen shot of IRSA PTF nightly summary page.

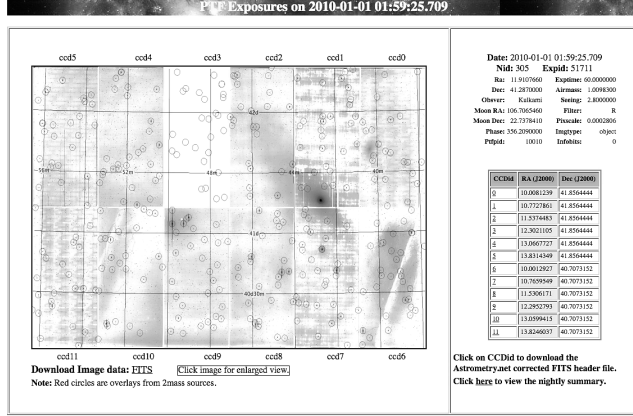


Figure 7. Screen shot of IRSA exposure preview, with 2MASS sources overlaid on the field.

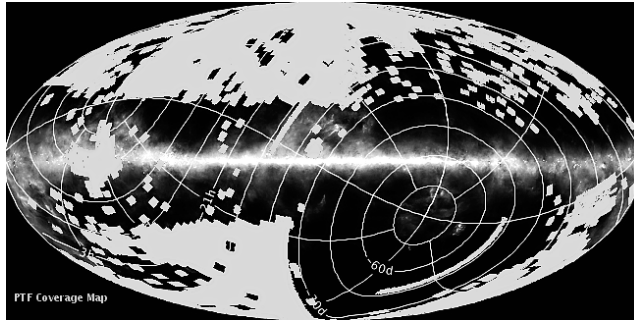


Figure 8. IRSA coverage map of all PTF image data taken as of 15 January, 2010.

8 Future Plans

We are currently studying the costs and methods associated with producing and serving:

- Co-added images
- Differential photometry
- Improved photometric calibration
- Variability classification and searching

Co-added images would enable detection of sources up to two magnitudes fainter than the current $m_R \approx 20.6$ limiting magnitude of individual images. Differential photometry would permit relative photometric accuracies approaching the stability limits of the detector. Depending on the level of accuracy achieved, such a capability may enable the routine detection of transiting substellar objects.

Over the course of the survey we expect to generate a very large catalog of stars outside the SDSS footprint which have been observed and well calibrated under photometric conditions. Using this catalog as a reference will ultimately enable us to observe the entire observable portion of the sky under less than ideal photometric conditions while still achieving our photometric accuracy goals.

On a somewhat longer time scale, we are studying automated methods of classifying variable sources using the Merged Sources table. Suitable metrics could then be computed, stored, and used to quickly search for variables of a particular type (e.g. RR Lyrae, cataclysmic variables, eclipsing binaries, etc.).

9 Conclusion

The IPAC PTF pipeline and archive is an efficient, cost-effective, and user-friendly approach to enabling a new and largely unexplored field of research, namely wide field, time domain astronomy. In addition to facilitating a large number of ongoing scientific investigations, the IPAC PTF pipeline is serving as a useful pathfinder for exploring the processing, handling, and serving of extremely large data sets. This overview provides little more than an outline of the pipeline and archive; a more detailed description of the design and operation of the system is given by Laher et al. (2010).

Acknowledgments. We are grateful to Nouhad Hamam, Eran Ofek, Nick Law, Robert Quimby, John Good, and Serge Monkenwitz for their many contributions to the design and development of the IPAC PTF pipeline.

References

- Bertin, E. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 112
- Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
- Laher, R. et al. 2010, in preparation
- Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S. 2009, arXiv:0910.2233
- Law, N. et al. 2009, PASP, 121, 1395
- Rau, A. et al. 2009, PASP, 121, 1334
- Skrutskie, M. F. et al. 2006, AJ, 131, 1163
- York, D. G. et al. 2000, AJ, 120, 1579